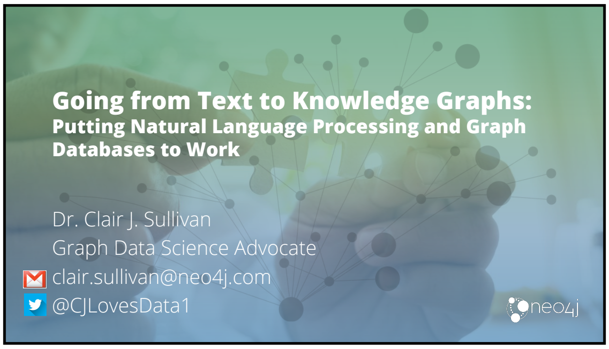
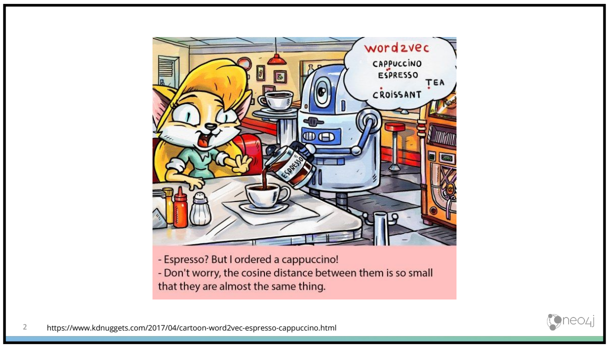


Going from Text to Knowledge Graphs:

Putting Natural Language Processing and Graph Databases to Work

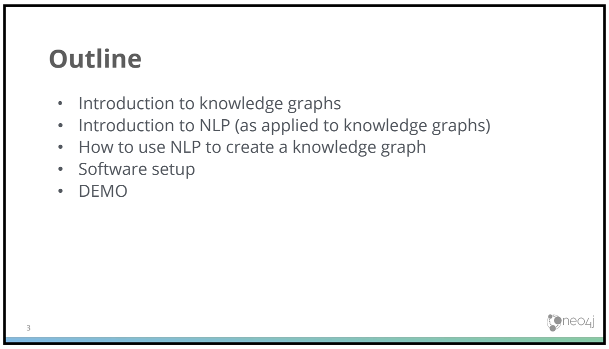
Dr. Clair J. Sullivan
Graph Data Science Advocate
clair.sullivan@neo4j.com
@CJLovesData1

neo4j



- Espresso? But I ordered a cappuccino!
- Don't worry, the cosine distance between them is so small that they are almost the same thing.

https://www.kdnuggets.com/2017/04/cartoon-word2vec-espresso-cappuccino.html

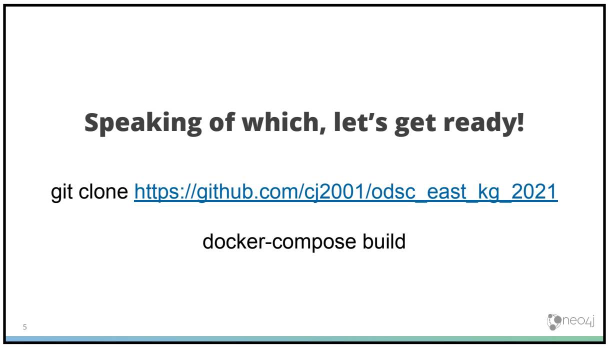


Outline

- Introduction to knowledge graphs
- Introduction to NLP (as applied to knowledge graphs)
- How to use NLP to create a knowledge graph
- Software setup
- DEMO

All materials (including the slides!) for this workshop are available on the workshop GitHub repo:

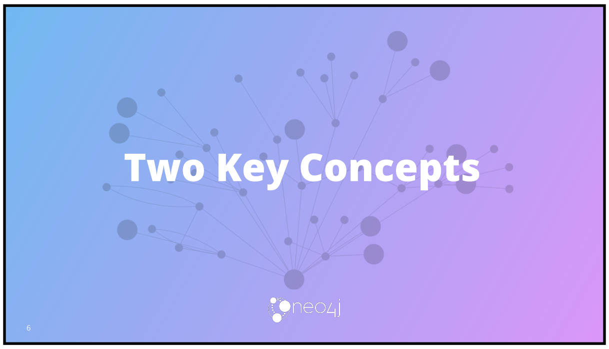
https://github.com/cj2001/odsc_east_kg_2021



Speaking of which, let's get ready!

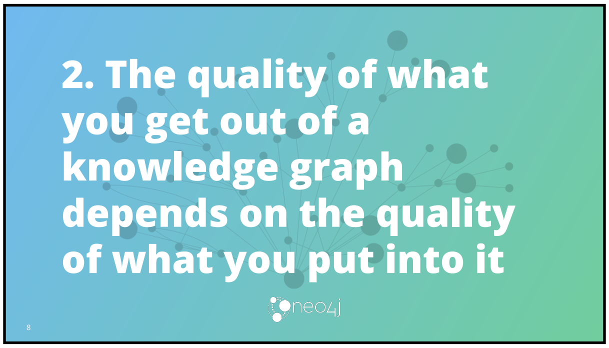
```
git clone https://github.com/cj2001/odsc_east_kg_2021

docker-compose build
```

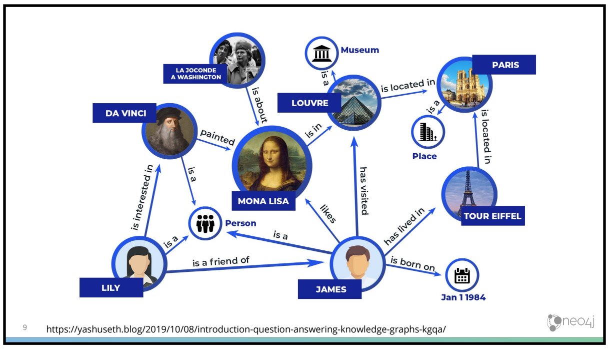


Two Key Concepts

1. There is no proverbial "silver bullet" with NLP



2. The quality of what you get out of a knowledge graph depends on the quality of what you put into it

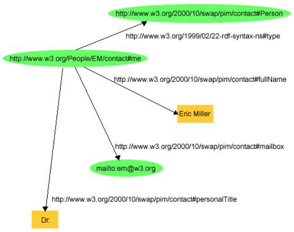


https://yashuseth.blog/2019/10/08/introduction-question-answering-knowledge-graphs-kgqa/

Introduction to knowledge graphs

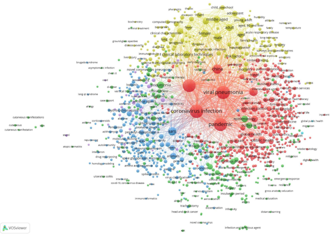
- “Things not strings”
- What knowledge graphs are useful for
  - Search
  - Question answering
  - Recommendation engine
- Can be generated a lot of different ways
  - Resource Description Framework (RDF)
  - Co-occurrence
  - Subject-Verb-Object (SVO)

RDF triples


http://www.w3.org/2000/10/swap/pim/contact#Person
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://www.w3.org/People/EM/contact#me
http://www.w3.org/2000/10/swap/pim/contact#fullName
Eric Miller
http://www.w3.org/2000/10/swap/pim/contact#mailbox
mailto:em@w3.org
http://www.w3.org/2000/10/swap/pim/contact#personalTitle
Dr.

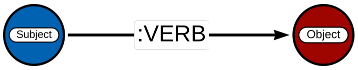
https://en.wikipedia.org/wiki/Resource_Description_Framework#Examples

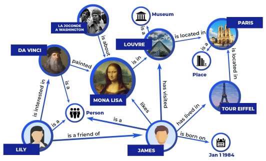
Word co-occurrence



https://covid19biblio.com/2020/04/28/keyword-co-occurrence-network-graph-for-the-overall-research-field-on-covid-19-up-to-april-27th-2020/

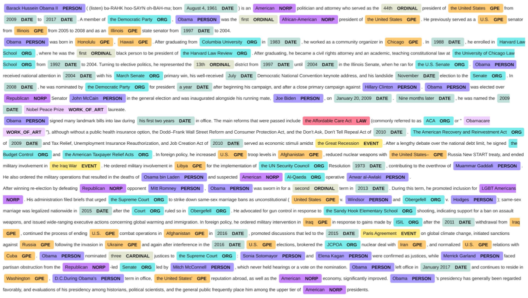
SVO triples





NLP considerations for knowledge graph creation

- Named Entity Recognition (NER)
- SVO / SPO triples
  - ...but verbs can be difficult, as you will see!
- Very language dependent
- Very topic-area dependent



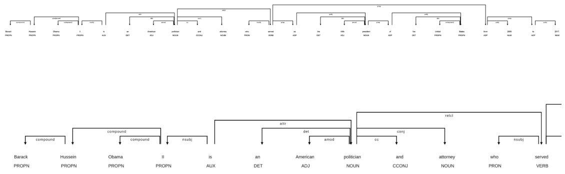
Barack Hussein Obama II PERSON ((listen) bə-RAHK hoo-SAYN oh-BAH-mə; born August 4, 1961 DATE) is an American NORP politician and attorney who served as the 44th ORDINAL president of the United States GPE from 2009 DATE to 2017 DATE . A member of the Democratic Party ORG , Obama PERSON was the first ORDINAL African-American NORP president of the United States GPE . He previously served as a U.S. GPE senator from Illinois GPE from 2005 to 2008 and as an Illinois GPE state senator from 1997 DATE to 2004.

Obama PERSON was born in Honolulu GPE , Hawaii GPE . After graduating from Columbia University ORG in 1983 DATE , he worked as a community organizer in Chicago GPE . In 1988 DATE , he enrolled in Harvard Law School ORG , where he was the first ORDINAL black person to be president of the Harvard Law Review ORG . After graduating, he became a civil rights attorney and an academic, teaching constitutional law at the University of Chicago Law School ORG from 1992 DATE to 2004. Turning to elective politics, he represented the 13th ORDINAL district from 1997 DATE until 2004 DATE in the Illinois Senate, when he ran for the U.S. Senate ORG . Obama PERSON received national attention in 2004 DATE with his March Senate ORG primary win, his well-received July DATE Democratic National Convention keynote address, and his landslide November DATE election to the Senate ORG . In 2008 DATE , he was nominated by the Democratic Party ORG for president a year DATE after beginning his campaign, and after a close primary campaign against Hillary Clinton PERSON . Obama PERSON was elected over Republican NORP Senator John McCain PERSON in the general election and was inaugurated alongside his running mate, Joe Biden PERSON , on January 20, 2009 DATE . Nine months later DATE , he was named the 2009 DATE Nobel Peace Prize WORK_OF_ART laureate.

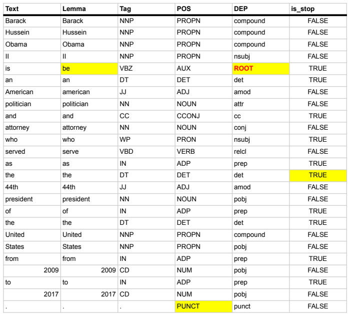
Obama PERSON signed many landmark bills into law during his first two years DATE in office. The main reforms that were passed include the Affordable Care Act LAW (commonly referred to as ACA ORG or " Obamacare WORK_OF_ART "), although without a public health insurance option, the Dodd–Frank Wall Street Reform and Consumer Protection Act, and the Don't Ask, Don't Tell Repeal Act of 2010 DATE . The American Recovery and Reinvestment Act ORG of 2009 DATE and Tax Relief, Unemployment Insurance Reauthorization, and Job Creation Act of 2010 DATE served as economic stimuli amidst the Great Recession EVENT . After a lengthy debate over the national debt limit, he signed the Budget Control ORG and the American Taxpayer Relief Acts ORG . In foreign policy, he increased U.S. GPE troop levels in Afghanistan GPE , reduced nuclear weapons with the United States– GPE Russia New START treaty, and ended military involvement in the Iraq War EVENT . He ordered military involvement in Libya GPE for the implementation of the UN Security Council ORG Resolution 1973 DATE , contributing to the overthrow of Muammar Gaddafi PERSON . He also ordered the military operations that resulted in the deaths of Osama bin Laden PERSON and suspected American NORP Al-Qaeda ORG operative Anwar al-Awlaki PERSON .

After winning re-election by defeating Republican NORP opponent Mitt Romney PERSON , Obama PERSON was sworn in for a second ORDINAL term in 2013 DATE . During this term, he promoted inclusion for LGBT Americans NORP . His administration filed briefs that urged the Supreme Court ORG to strike down same-sex marriage bans as unconstitutional (United States GPE v. Windsor PERSON and Obergefell ORG v. Hodges PERSON); same-sex marriage was legalized nationwide in 2015 DATE after the Court ORG ruled so in Obergefell ORG . He advocated for gun control in response to the Sandy Hook Elementary School ORG shooting, indicating support for a ban on assault weapons, and issued wide-ranging executive actions concerning global warming and immigration. In foreign policy, he ordered military intervention in Iraq GPE in response to gains made by ISIL ORG after the 2011 DATE withdrawal from Iraq GPE , continued the process of ending U.S. GPE combat operations in Afghanistan GPE in 2016 DATE , promoted discussions that led to the 2015 DATE Paris Agreement EVENT on global climate change, initiated sanctions against Russia GPE following the invasion in Ukraine GPE and again after interference in the 2016 DATE U.S. GPE elections, brokered the JCPOA ORG nuclear deal with Iran GPE , and normalized U.S. GPE relations with Cuba GPE . Obama PERSON nominated three CARDINAL justices to the Supreme Court ORG : Sonia Sotomayor PERSON and Elena Kagan PERSON were confirmed as justices, while Merrick Garland PERSON faced partisan obstruction from the Republican NORP -led Senate ORG led by Mitch McConnell PERSON , which never held hearings or a vote on the nomination. Obama PERSON left office in January 2017 DATE and continues to reside in Washington GPE , D.C.During Obama's PERSON term in office, the United States' GPE reputation abroad, as well as the American NORP economy, significantly improved. Obama PERSON 's presidency has generally been regarded favorably, and evaluations of his presidency among historians, political scientists, and the general public frequently place him among the upper tier of American NORP presidents.

Barack Hussein Obama II is an American politician and attorney who served as the 44th president of the United States from 2009 to 2017.

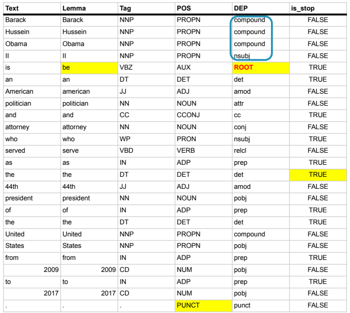


Barack Hussein Obama II is an American politician and attorney who served as the 44th president of the United States from 2009 to 2017.



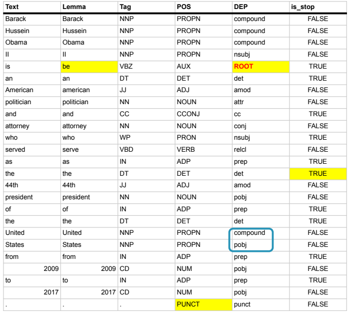
Text	Lemma	Tag	POS	DEP	is_stop
Barack	Barack	NNP	PROPN	compound	FALSE
Hussein	Hussein	NNP	PROPN	compound	FALSE
Obama	Obama	NNP	PROPN	compound	FALSE
II	II	NNP	PROPN	nsubj	FALSE
is	be	VBZ	AUX	ROOT	TRUE
an	an	DT	DET	det	TRUE
American	american	JJ	ADJ	amod	FALSE
politician	politician	NN	NOUN	attr	FALSE
and	and	CC	CCONJ	cc	TRUE
attorney	attorney	NN	NOUN	conj	FALSE
who	who	WP	PRON	nsubj	TRUE
served	serve	VBD	VERB	relcl	FALSE
as	as	IN	ADP	prep	TRUE
the	the	DT	DET	det	TRUE
44th	44th	JJ	ADJ	amod	FALSE
president	president	NN	NOUN	pobj	FALSE
of	of	IN	ADP	prep	TRUE
the	the	DT	DET	det	TRUE
United	United	NNP	PROPN	compound	FALSE
States	States	NNP	PROPN	pobj	FALSE
from	from	IN	ADP	prep	TRUE
2009	2009	CD	NUM	pobj	FALSE
to	to	IN	ADP	prep	TRUE
2017	2017	CD	NUM	pobj	FALSE
.	.	.	PUNCT	punct	FALSE

Barack Hussein Obama II is an American politician and attorney who served as the 44th president of the United States from 2009 to 2017.



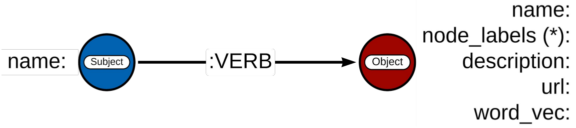
Text	Lemma	Tag	POS	DEP	is_stop
Barack	Barack	NNP	PROPN	compound	FALSE
Hussein	Hussein	NNP	PROPN	compound	FALSE
Obama	Obama	NNP	PROPN	compound	FALSE
II	II	NNP	PROPN	nsubj	FALSE
is	be	VBZ	AUX	ROOT	TRUE
an	an	DT	DET	det	TRUE
American	american	JJ	ADJ	amod	FALSE
politician	politician	NN	NOUN	attr	FALSE
and	and	CC	CCONJ	cc	TRUE
attorney	attorney	NN	NOUN	conj	FALSE
who	who	WP	PRON	nsubj	TRUE
served	serve	VBD	VERB	relcl	FALSE
as	as	IN	ADP	prep	TRUE
the	the	DT	DET	det	TRUE
44th	44th	JJ	ADJ	amod	FALSE
president	president	NN	NOUN	pobj	FALSE
of	of	IN	ADP	prep	TRUE
the	the	DT	DET	det	TRUE
United	United	NNP	PROPN	compound	FALSE
States	States	NNP	PROPN	pobj	FALSE
from	from	IN	ADP	prep	TRUE
2009	2009	CD	NUM	pobj	FALSE
to	to	IN	ADP	prep	TRUE
2017	2017	CD	NUM	pobj	FALSE
.	.	.	PUNCT	punct	FALSE

Barack Hussein Obama II is an American politician and attorney who served as the 44th president of the United States from 2009 to 2017.



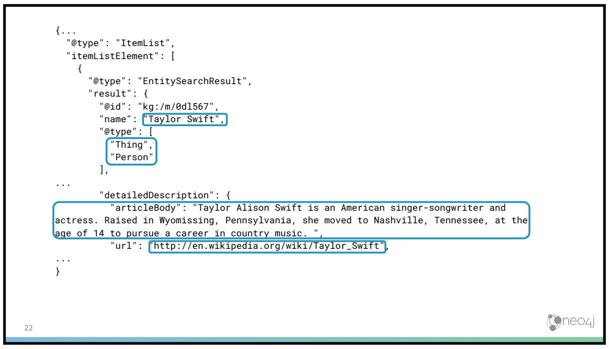
Text	Lemma	Tag	POS	DEP	is_stop
Barack	Barack	NNP	PROPN	compound	FALSE
Hussein	Hussein	NNP	PROPN	compound	FALSE
Obama	Obama	NNP	PROPN	compound	FALSE
II	II	NNP	PROPN	nsubj	FALSE
is	be	VBZ	AUX	ROOT	TRUE
an	an	DT	DET	det	TRUE
American	american	JJ	ADJ	amod	FALSE
politician	politician	NN	NOUN	attr	FALSE
and	and	CC	CCONJ	cc	TRUE
attorney	attorney	NN	NOUN	conj	FALSE
who	who	WP	PRON	nsubj	TRUE
served	serve	VBD	VERB	relcl	FALSE
as	as	IN	ADP	prep	TRUE
the	the	DT	DET	det	TRUE
44th	44th	JJ	ADJ	amod	FALSE
president	president	NN	NOUN	pobj	FALSE
of	of	IN	ADP	prep	TRUE
the	the	DT	DET	det	TRUE
United	United	NNP	PROPN	compound	FALSE
States	States	NNP	PROPN	pobj	FALSE
from	from	IN	ADP	prep	TRUE
2009	2009	CD	NUM	pobj	FALSE
to	to	IN	ADP	prep	TRUE
2017	2017	CD	NUM	pobj	FALSE
.	.	.	PUNCT	punct	FALSE

Detailed knowledge graph data model

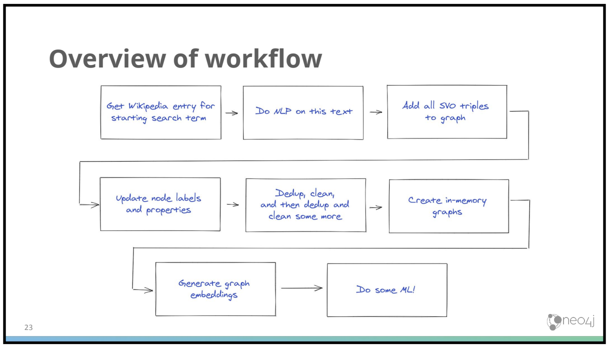


An introduction to the tools we will use today

- spacy
- Wikipedia Python package
- Google Knowledge Graph
- Neo4j
  - Awesome Procedures on Cypher (APOC)
  - Graph Data Science (GDS) Library
  - Cypher
    - No Cypher knowledge is assumed!



```
{...
  "@type": "ItemList",
  "itemListElement": [
    {
      "@type": "EntitySearchResult",
      "result": {
        "@id": "kg:/m/0dl567",
        "name": "Taylor Swift",
        "@type": [
          "Thing",
          "Person"
        ],
...
        "detailedDescription": {
          "articleBody": "Taylor Alison Swift is an American singer-songwriter and
actress. Raised in Wyomissing, Pennsylvania, she moved to Nashville, Tennessee, at the
age of 14 to pursue a career in country music. ",
          "url": "http://en.wikipedia.org/wiki/Taylor_Swift",
...
}
```



Overview of workflow

Get Wikipedia entry for starting search term

Do NLP on this text

Add all SVO triples to graph

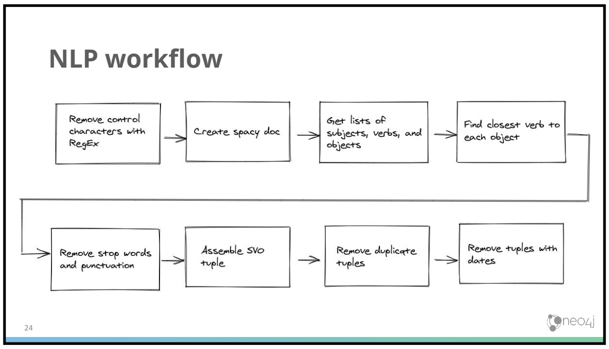
Update node labels and properties

Dedup, clean, and then dedup and clean some more

Create in-memory graphs

Generate graph embeddings

Do some ML!



NLP workflow

Remove control characters with RegEx

Create spacy doc

Get lists of subjects, verbs, and objects

Find closest verb to each object

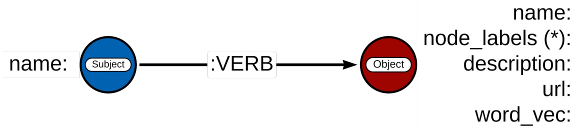
Remove stop words and punctuation

Assemble SVO tuple

Remove duplicate tuples

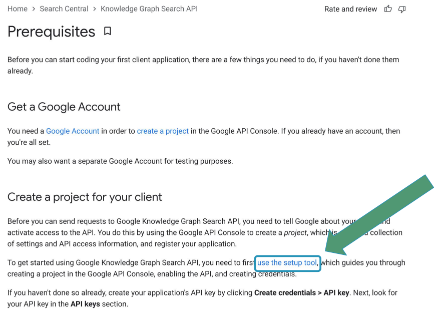
Remove tuples with dates

Detailed knowledge graph data model



Software setup

- You will need access to Docker, CLI, and a browser
- Clone the GitHub repository:
  - https://github.com/cj2001/odsc_east_kg_2021
- Google KG API key
  - https://developers.google.com/knowledge-graph/prereqs
  - Get this key and then save it in /notebooks/.api_key



Prerequisites

Before you can start coding your first client application, there are a few things you need to do, if you haven't done them already.

Get a Google Account

You need a Google Account in order to create a project in the Google API Console. If you already have an account, then you're all set.

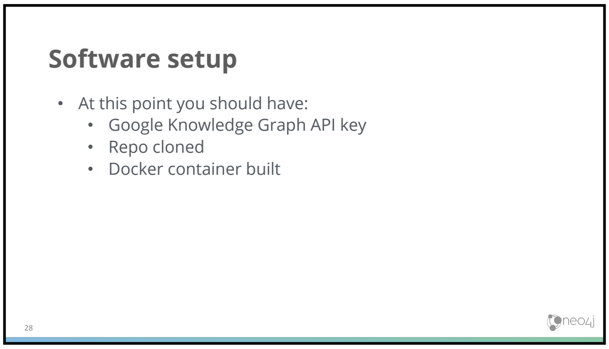
You may also want a separate Google Account for testing purposes.

Create a project for your client

Before you can send requests to Google Knowledge Graph Search API, you need to tell Google about your [illegible] and activate access to the API. You do this by using the Google API Console to create a *project*, which is [illegible] collection of settings and API access information, and register your application.

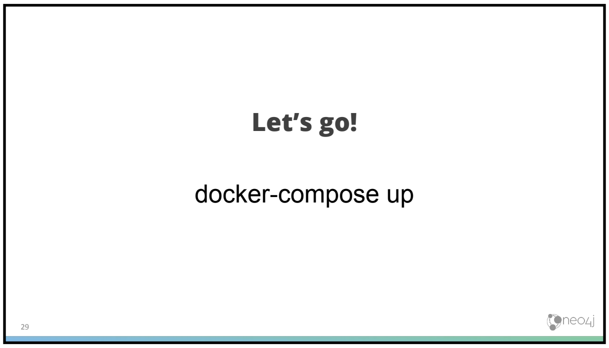
To get started using Google Knowledge Graph Search API, you need to first use the setup tool, which guides you through creating a project in the Google API Console, enabling the API, and creating credentials.

If you haven't done so already, create your application's API key by clicking **Create credentials > API key**. Next, look for your API key in the **API keys** section.



Software setup

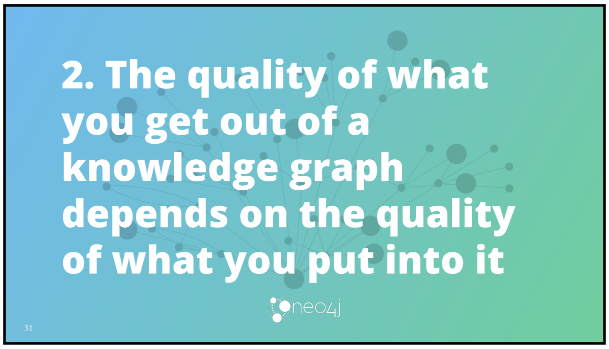
- At this point you should have:
  - Google Knowledge Graph API key
  - Repo cloned
  - Docker container built



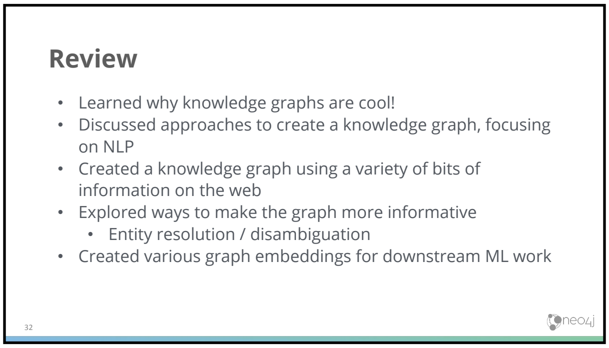
Let's go!

docker-compose up

1. There is no proverbial "silver bullet" with NLP

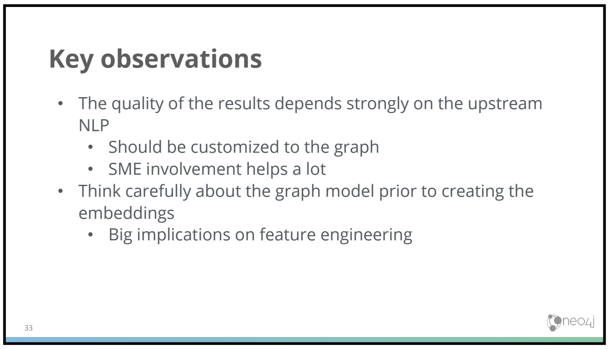


2. The quality of what you get out of a knowledge graph depends on the quality of what you put into it



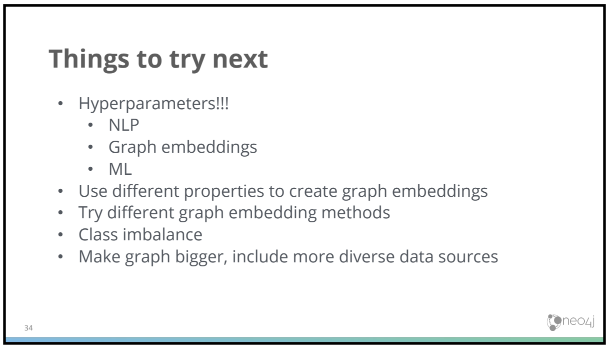
Review

- Learned why knowledge graphs are cool!
- Discussed approaches to create a knowledge graph, focusing on NLP
- Created a knowledge graph using a variety of bits of information on the web
- Explored ways to make the graph more informative
  - Entity resolution / disambiguation
- Created various graph embeddings for downstream ML work



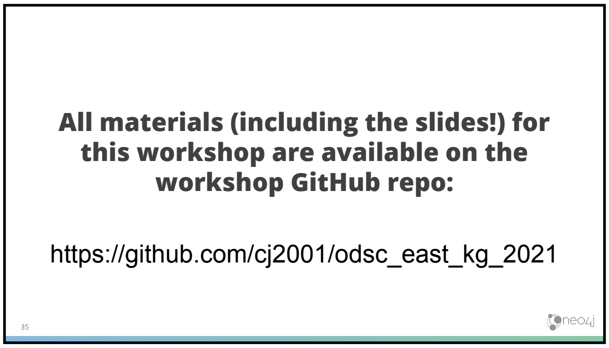
Key observations

- The quality of the results depends strongly on the upstream NLP
  - Should be customized to the graph
  - SME involvement helps a lot
- Think carefully about the graph model prior to creating the embeddings
  - Big implications on feature engineering



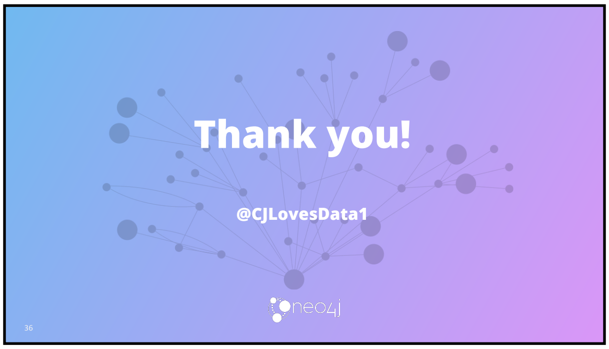
Things to try next

- Hyperparameters!!!
  - NLP
  - Graph embeddings
  - ML
- Use different properties to create graph embeddings
- Try different graph embedding methods
- Class imbalance
- Make graph bigger, include more diverse data sources



All materials (including the slides!) for this workshop are available on the workshop GitHub repo:

https://github.com/cj2001/odsc_east_kg_2021



Thank you!

@CJLovesData1